Speaker Segmentation on Conversational Telephone Speech

Qin Jin
Interactive Systems Laboratory
Carnegie Mellon University

NIST RT03 Workshop
19-20 May 2003

Outline

- Segmentation with speech only
  - **Speech Detection: split the conversation from each side into speech and silence/noise**
    1. Raw segmentation
    2. Model training
    3. Resegmentation
    4. Raw smoothing
    5. Iteration of 2, 3 and 4
    6. Final segmentation and final smoothing
  - **Gender Recognition**
- Segmentation with reference
- Conclusions and Discussions

segmentation with speech only

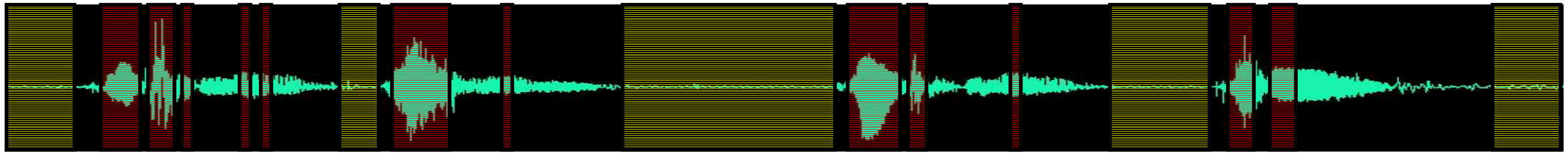
raw segmentation

- Classify the speech signal into three classes: highly-confident speech, highly-confident silence, unsure
  - **Frame size 30ms, window shift 10ms**
  - **Decision Criteria**
    - Energy
    - Zero-crossing rate
    - FFT magnitude variance (a speech frame has higher variance than a silence frame)
  - **Separate thresholds for speech and silence**
  - **Decision**
    - Highly-confident speech frame
    - Highly-confident silence frame
    - Unsure frame

segmentation with speech only

model training, resegmentation and smoothing

- **Step2: Train GMMs on highly-confident speech and silence frames**
- **Step 3: Classify the unsure frames using trained GMMs**
  - If P(x|GMM-spch) / P(x|GMM-sil) > TH then “speech frame”
  - If P(x|GMM-sil) / P(x|GMM-spch) > TH then “silence frame”
  - Otherwise “unsure frame”
- **Step 4: Smooth out potential speech segments or silence segments via segment-length threshold**
  - Speech >= 0.2s, Silence>=0.1s, otherwise re-label the segment as “unsure”
- **Step 5: Iterate the model training, resegmentation and smoothing several times (final system: 5 iterations)**



GMM-spch
GMM-sil

segmentation with speech only

Final segmentation and smoothing

- Classify the remained “unsure” frames as either speech or silence according to P(x|GMM)
- Final smoothing via different segment-length threshold
  - **Speech >= 0.05s, Silence >= 0.03s**
  - **If a segment doesn’t satisfy the criteria then merge it to its neighbor segment**
    - Final system : left neighbor
  - **RT03 evaluation results: miss 9.1%, false alarm: 2.3%**

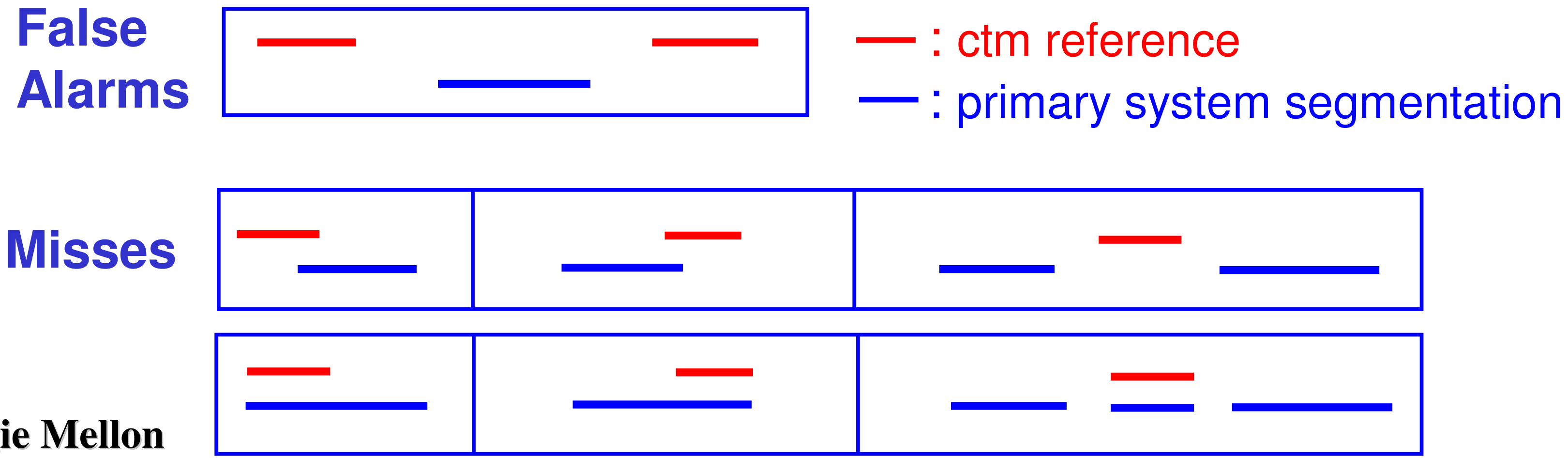
segmentation with speech only

Gender recognition

- Independent step from segmentation
- Gender identity is decided based on the speech segments only
- Adult-female and adult-male GMMs are trained using randomly chosen conversations of hub5e_01 dataset
  - **Data: ~60m for each gender**
    - Balanced gender distribution
    - All acoustic conditions (swb1, swb2, swb_cell)
  - **Features: 20 cepstral coefficients**
  - **Models: 256 Mixtures of Gaussians**

segmentation with reference

- First step: Segmentation with speech data only (primary system)
- Refine the segmentation according to ctm reference
- Merge two segments if the pause between them is less than 0.3 seconds
- RT03 evaluation results: miss 1.1%, false alarm: 1.7%



Conclusions and Discussions

- speech segmentation on conversational telephone speech using speech data only
  - **Unsupervised adaptation for segmentation**
  - **System can be applied for other data of different acoustic conditions with no change**
- Speech segmentation with reference data
  - **A straight forward approach**
- Discussions
  - **Does the “segmentation with reference” task make sense?**
  - **More efficient and cooperative approach for segmentation with reference**